Contribution ID: **32** Type: **Long Talk**

Management of Open Data Lifecycle with Onedata

Tuesday, 1 October 2024 15:35 (15 minutes)

Onedata[1] is a high-performance data management system with a distributed, global infrastructure that enables users to access heterogeneous storage resources worldwide. It supports various use cases ranging from personal data management to data-intensive scientific computations. Onedata has a fully distributed architecture that facilitates the creation of a hybrid cloud infrastructure with private and commercial cloud resources. Users can collaborate, share, and publish data, as well as perform high-performance computations on distributed data using different interfaces: POSIX-compliant native mounts, pyfs (python filesystem) plugins, REST/CDMI API, and S3 protocol (currently in beta).

The latest Onedata release line, 21.02, introduces several new features and improvements that enhance its capabilities in managing distributed datasets throughout their lifecycle. The software allows users to establish a hierarchical structure of datasets, control multi-site replication and distribution using Quality-of-Service rules, and keep track of the dataset size statistics over time. In addition, it also supports the annotation of datasets with metadata, which is crucial for organising and searching for specific data. The platform also includes robust protection mechanisms that prevent data and metadata modification, ensuring the integrity of the dataset in its final stage of preparation. Another key feature of Onedata is its ability to archive datasets for long-term preservation, enabling organisations to retain critical data for future use. This is especially useful in fields such as scientific research, where datasets are often used for extended periods or cited in academic papers. Finally, Onedata supports data-sharing mechanisms aligned with the idea of Open Data, such as the OAI-PMH protocol and the newly introduced Space Marketplace. These features enable users to easily share their datasets with others, either openly or through controlled access.

Currently, Onedata is used in European projects: EUreka3D[2], EuroScienceGateway[3], DOME[4], and InterTwin[5], where it provides a data transparency layer for managing large, distributed datasets on dynamic hybrid cloud containerised environments.

Acknowledgements: This work is co-financed by the Polish Ministry of Education and Science under the program entitled International Co-financed Projects (projects no. 5398/DIGITAL/2023/2 and 5399/DIGITAL/2023/2).

REFERENCES:
- Onedata project website. https://onedata.org. EUreka3D: European Union's REKonstructed in 3D. https://eureka3d.eu. EuroScienceGateway project: open infrastructure for data-driven research. https://galaxyproject.org/projects/esg/. DOME: A Distributed Open Marketplace for Europe Cloud and Edge Services. https://dome-marketplace.eu. InterTwin: Interdisciplinary Digital Twin Engine for Science. https://intertwin.eu.

Topic

Data innovations: Data Management/Integration/Exchange

Primary authors: KRYZA, Bartosz (CYFRONET); DUTKA, Lukasz (CYFRONET); OPIOLA, Lukasz (CYFRONET); ORZECHOWSKI, Michal (CYFRONET)

Presenter: ORZECHOWSKI, Michal (CYFRONET)

Session Classification: Inside Data Spaces: Enabling data sharing paradigms